

(19) World Intellectual Property
Organization
International Bureau



(43) International Publication Date
3 February 2005 (03.02.2005)

PCT

(10) International Publication Number
WO 2005/010727 A2

- (51) International Patent Classification⁷: **G06F**
- (21) International Application Number:
PCT/US2004/023932
- (22) International Filing Date: 23 July 2004 (23.07.2004)
- (25) Filing Language: English
- (26) Publication Language: English
- (30) Priority Data:
60/489,454 23 July 2003 (23.07.2003) US
- (71) Applicant (for all designated States except US):
PRAEDEA SOLUTIONS, INC. [US/US]; Suite 1008,
131 Varick Street, New York, NY 10013 (US).
- (72) Inventors; and
- (75) Inventors/Applicants (for US only): **GRAF, James, A.** [US/US]; Suite 1008, 131 Varick Street, New York, NY 10013 (US). **KOROTEYEV, Vladimir, A.** [US/US]; #3A West, 8801 Shore Road, Brooklyn, NY 11209 (US). **MIKHAYLOV, Eduard, Y.** [US/US]; 108-28 65th Road, Forest Hills, NY 11375 (US). **BRICKER, Elliot, I.**

[US/US]; #8C, 711 Amsterdam Avenue, New York, NY 10025 (US). **LEVY, Benjamin, D., A.** [FR/US]; 375 Faxon Avenue, San Francisco, CA 94112 (US). **WONG, Augustinus, Y.** [US/US]; 99 Sussex Road, Tenafly, NJ 07670 (US).

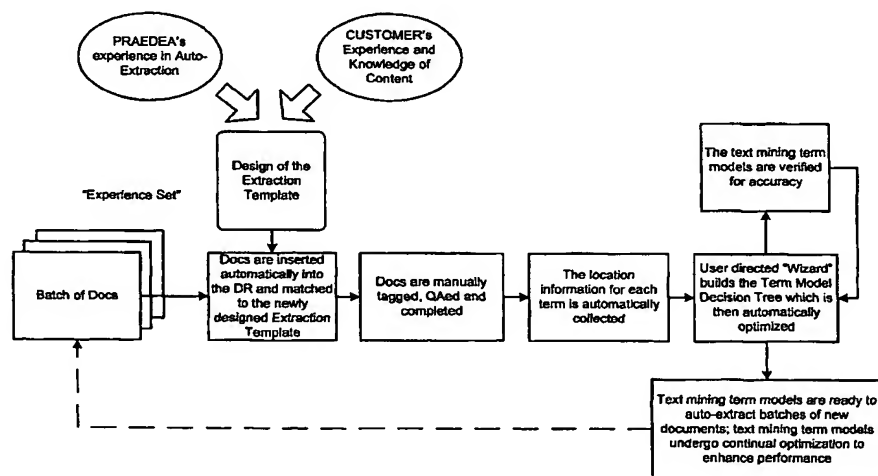
(74) Agent: **FORREST, Peter**; Gray, Plant, Mooty, Mooty & Bennett, P.A., P.O. Box 2906, Minneapolis, MN 55402-0906 (US).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM),

[Continued on next page]

(54) Title: EXTRACTING DATA FROM SEMI-STRUCTURED TEXT DOCUMENTS





European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— without international search report and to be republished upon receipt of that report

Declaration under Rule 4.17:

— of inventorship (Rule 4.17(iv)) for US only

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.